(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2023-537193
(P2023-537193A)
(43)公表日　令和5年8月31日(2023.8.31)

(51)Int.Cl.	F I	テーマコード(参考)
G06F 16/906 (2019.01)	G06F 16/906	5B175

審査請求 有　予備審査請求 未請求　(全 15 頁)

(21)出願番号　特願2022-578769(P2022-578769)
(86)(22)出願日　令和4年3月15日(2022.3.15)
(85)翻訳文提出日　令和5年2月13日(2023.2.13)
(86)国際出願番号　PCT/IB2022/052333
(87)国際公開番号　WO2022/269368
(87)国際公開日　令和4年12月29日(2022.12.29)
(31)優先権主張番号　202141028706
(32)優先日　令和3年6月25日(2021.6.25)
(33)優先権主張国・地域又は機関
インド(IN)

(71)出願人　520412811
エルアンドティー　テクノロジー　サービシズ　リミテッド
インド国，チェンナイ　６０００８９，ラーマプラーム，マウント　プーナマリー　ロード，ディーエルエフ　アイティー　エスイーゼット　パーク，ブロック　３，２番　フロア，１／１２４
(74)代理人　110002365
弁理士法人サンネクスト国際特許事務所
(72)発明者　ダス，　イシタ
インド　７００００８　ウエストベンガル，　コルカタ，　イーストーパーク，　バリシャ，　カリパダ　ムカルジー　ロード，　３３９／ジー／１７

最終頁に続く

(54)【発明の名称】クラスタを表現するためにサンプルを選択する方法およびシステム

(57)【要約】
クラスタを表現するためにサンプルを選択する方法が開示される。方法は、最適化デバイスによって、１つまたは複数のクラスタを受信することを含んでもよい。１つまたは複数のクラスタのそれぞれは、複数のサンプルを含む。方法は、１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントを決定してもよく、１つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列を生成してもよい。方法は、クラスタの複数のサンプルを、クラスタ内の複数のサンプルのばらつきの度合いに基づいて、ソートしてもよい。ソートすることは、１つまたは複数のクラスタのそれぞれについてのアレイベースの距離行列を使用して実施してもよい。さらに、方法は、クラスタを表現するために、サンプルの数の決定されたカウントを、複数のクラスタのそれぞれの、ソートされた複数のサンプルから選択してもよい。
【選択図】図３

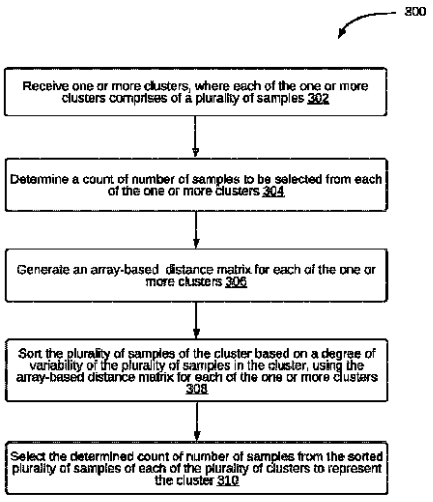

FIG. 3

【特許請求の範囲】
【請求項１】
クラスタを表現するためにサンプルを選択する方法であって、
最適化デバイスによって、それぞれが複数のサンプルを備える１つまたは複数のクラスタを受信し、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントを決定し、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列を生成し、
前記最適化デバイスによって、前記クラスタの前記複数のサンプルを、前記クラスタ内の前記複数のサンプルのばらつきの度合いに基づいて、前記１つまたは複数のクラスタのそれぞれについての前記アレイベースの距離行列を使用して、ソートし、
前記最適化デバイスによって、前記クラスタを表現するために、サンプルの数の前記決定されたカウントを、前記複数のクラスタのそれぞれの、前記ソートされた複数のサンプルから選択する、
方法。
【請求項２】
前記サンプルの数の前記選択された決定されたカウントがしきい値未満である場合、前記サンプルの数の所定のカウントを、前記１つまたは複数のクラスタのそれぞれから選択し、前記しきい値が、各データセットに特有であり、クラスタのサイズを前記データセットと比較することによって決定される、
請求項１に記載の方法。
【請求項３】
前記１つまたは複数のクラスタのそれぞれから選択されるサンプルの数の前記カウントが、前記１つまたは複数のクラスタのそれぞれについてのサイズ、ばらつき、およびクラスタ確率のうちの少なくとも１つに基づいて、層化サンプリング技法を使用して、決定される、
請求項１に記載の方法。
【請求項４】
前記クラスタ確率が、機械学習（ＭＬ）モデルを使用して決定され、前記ＭＬモデルが、前記クラスタの前記複数のサンプルを分類する、
請求項３に記載の方法。
【請求項５】
下記を行なうように構成された１つまたは複数のコンピューティングデバイスを備える、
最適化デバイスによって、それぞれが複数のサンプルを備える１つまたは複数のクラスタを受信する、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントを決定する、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列を生成する、
前記最適化デバイスによって、前記クラスタの前記複数のサンプルを、前記クラスタ内の前記複数のサンプルのばらつきの度合いに基づいて、前記１つまたは複数のクラスタのそれぞれについての前記アレイベースの距離行列を使用して、ソートする、および、
前記最適化デバイスによって、前記クラスタを表現するために、サンプルの数の前記決定されたカウントを、前記複数のクラスタのそれぞれの、前記ソートされた複数のサンプルから選択する、
システム。
【請求項６】
前記サンプルの数の前記選択された決定されたカウントがしきい値未満である場合、前記サンプルの数の所定のカウントを、前記１つまたは複数のクラスタのそれぞれから選択し

、前記しきい値が、各データセットに特有であり、クラスタのサイズを前記データセットと比較することによって決定される、
請求項５に記載のシステム。
【請求項７】
前記１つまたは複数のクラスタのそれぞれから選択されるサンプルの数の前記カウントが、１つまたは複数のクラスタのそれぞれについてのサイズ、ばらつき、およびクラスタ確率のうちの少なくとも１つに基づいて、層化サンプリング技法を使用して、決定される、
請求項５に記載のシステム。
【請求項８】
前記クラスタ確率が、機械学習（ＭＬ）モデルを使用して決定され、前記ＭＬモデルが、前記クラスタの前記複数のサンプルを分類する、
請求項７に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、一般に、データセットのサイズを小さくすることに関し、より詳細には、データセットのサイズを小さくするためのクラスタを表現するために複数のサンプルを選択することに関する。
【背景技術】
【０００２】
情報爆発を伴うビッグデータの時代では、データ処理に関連するビジネス要件が、日々高まっており、処理対象のデータサンプルは、益々複雑化している。データのクラスタリングは、システムのエンジニアリングおよび計算科学において極めて重要な分野であり、教師無し学習を使用して、ラベル付けされたデータを伴わずに機械学習モデルを訓練する。クラスタリングアルゴリズムは、自身の一意な特徴を有するいくつかのクラスタに、データを分割する。時には、これらのクラスタ自身が、膨大な数のサンプルを有する。ラベル付けされていないデータが利用可能となるのは、高い次元、また線形に分離不可能なデータ空間であり、そのことにより、機械学習モデルの処理および訓練の間、大きなメモリチャンクおよび時間を消費する。
【発明の概要】
【０００３】
したがって、当分野では、機械学習モデルの効果的でリソース効率の良い訓練のために、クラスタ内のサンプルの総数のカウントを低減することによって、データセットのサイズを小さくするための方法およびシステムを提供する必要性がある。
【０００４】
クラスタを表現するためにサンプルを選択する方法が開示される。方法は、最適化デバイスによって、１つまたは複数のクラスタを受信することを含んでもよい。１つまたは複数のクラスタのそれぞれは、複数のサンプルを含む。方法は、１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントを決定してもよく、１つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列を生成してもよい。方法は、クラスタの複数のサンプルを、クラスタ内の複数のサンプルのばらつきの度合いに基づいて、ソートしてもよい。ソートすることは、１つまたは複数のクラスタのそれぞれについてのアレイベースの距離行列を使用して実施してもよい。さらに、方法は、クラスタを表現するために、サンプルの数の決定されたカウントを、複数のクラスタのそれぞれの、ソートされた複数のサンプルから選択してもよい。
【０００５】
本開示に組み込まれ、その一部を成す添付の図面は、例示の実施形態を図示しており、説明と併せて、開示される原理を説明するよう機能する。
【図面の簡単な説明】
【０００６】

【図１】本開示の実施形態による、１つまたは複数のクラスタから複数のデータサンプルを選択するためのプロセス図である。
【図２】本開示のいくつかの実施形態による、１つまたは複数のクラスタから複数のデータサンプルをソートして選択するためのプロセス図である。
【図３】本開示のいくつかの実施形態による、クラスタを表現するためにサンプルを選択する方法のフローチャートである。
【発明を実施するための形態】
【０００７】
例示的な実施形態を、添付の図面を参照して説明する。同一または類似の部分を参照するために、好都合であればいつでも、図面を通じて同一の参照符号を使用する。本明細書では開示される原理の例および特徴が説明されるが、開示される実施形態の思想および範囲から逸脱することなく、修正、適合、および他の実装が可能である。以下の詳細な説明は、単なる例として考えられ、真の範囲および思想は、後に続く特許請求の範囲によって示されることが意図されている。追加的な例示の実施形態を以降に列挙する。
【０００８】
理解されるように、クラスタリングアルゴリズムは、自身の一意な特徴を有するいくつかのクラスタに、データを分割する。時には、これらのクラスタ自身が、膨大な数のサンプルを有する。本開示は、クラスタに固有なばらつき、性質をカバーする、限られた数のサンプルを用いてクラスタを表現することができる解決策を提供する。このやり方で、アルゴリズムは、さらなるプロセスのためにデータセット全体を使用するために依存性を低減し、それにより、大きなデータセットを用いて作業する際のメモリおよび時間の複雑さを制限する。アルゴリズムはまた、クラスタのサイズが小さい場合に、ユーザが必要なサンプル数をクラスタから選択できるようにする、柔軟なものである。
【０００９】
さらには、プロセスは、均質なクラスタからでも、一意なサンプルを選択することができるよう確実にする。
【００１０】
図１を参照すると、本開示の実施形態による、１つまたは複数のクラスタから複数のデータサンプルを選択するためのプロセス１００が図示される。
【００１１】
ステップ１０２では、データセットは、１つまたは複数の異なるクラスタにクラスタ化されてもよい。クラスタリングは、同じように見え、類似の特徴を有するデータセットが、特定のクラスタ内に一緒に保持されることを確実にするよう実施されてもよい。
【００１２】
ステップ１０４では、複数のデータサンプルのうち、いくつのデータサンプルを、１つまたは複数の異なる作成済クラスタのうちの１つのクラスタから選択することができるかを決定することができる。
【００１３】
いくつのデータサンプルが選択されるかを決定するために、最適な割り当てのための、層化サンプリングメカニズムが使用されてもよい。層化サンプリングメカニズムは、複数のデータサンプルを考慮することができる。複数のデータサンプルのそれぞれは、均質なグループ（つまり、類似の特徴を有する複数のデータサンプルのそれぞれを一緒に格納することができる、クラスタ）に分割することができる。決定は、複数の類似して見えるサンプルの中から、いくつのサンプルが選択され得るかに関する場合がある。ステップ１０６では、データサンプルのうち、どれを、１つまたは複数の異なるクラスタから選択することができるかを決定することができる。層化サンプリングメカニズムは、特定の均質なデータグループのうちの１つを選択してもよく、特定の計算に基づいて１つまたは複数のデータサンプルをランダムに選択してもよい。
【００１４】
ステップ１０８では、所与のクラスタについて、下で言及する式を用いて、N_i個のサン

プルをクラスタから選択することができる：
【数１】

$$n_i = \frac{w_i s_i}{\sqrt{c_i}} / \frac{c_0}{\sum_{i=1}^{k} \sqrt{c_i} w_i s_i} / \quad \cdots \quad 式(1)$$

ただし、w_iは、ｉ番目のクラスタ内に存在するデータサンプルの数であり、
S_iは、クラスタ内のデータサンプルのばらつきであり、
C_iは、平均のクラスタ確率であり、
C_0は、定数である。
【００１５】
式（１）は、式（２）のように、簡単にすることができる。
【数２】

$$n = \sum_{i=1}^{k} n_i \quad \cdots \quad 式(2)$$

【００１６】
式（２）では、クラスタのサイズとばらつきを考慮してもよい。
【００１７】
クラスタ内に存在する複数のデータサンプル同士を区別するために機械学習モデルが利用可能であり、また所与のセットの特徴に基づいて複数のデータサンプル同士の区別の正確なレベルを予測することができる事例では、機械学習モデルから導かれる学習を利用して、クラスタ確率を考慮することによって、いくつのデータサンプルをクラスタから選択することができるかを決定することができる。区別のレベルに関する学習が、利用可能な機械学習モデルのいずれからも利用可能ではない場合、クラスタ確率パラメータは、データサンプルの数の選択を決定するための重み付けが、あまり与えられない場合がある。
【００１８】
ステップ１０６では、どのデータサンプルが選択されるかに関する決定が実施されてもよい。一般的に、データサンプルは、利用可能なランダム選択メカニズムのいずれかに基づいて、ランダムに選択され得る。しかしながら、選択されたデータサンプルがクラスタ全体を表現するべく、クラスタのばらつきの度合いを最大化するために、距離ベースの選択メカニズムが、ステップ１１０で利用されてもよい。距離ベースの選択メカニズムでは、クラスタ内に存在するアレイベースで最適化された距離が利用されてもよい。距離行列は、例えば、ユークリッドベースの距離行列またはマンハッタンベースの距離行列であってもよい。さらに、データサンプルは、その距離に基づいて、つまり、ばらつきの最大化に基づいてソートされてもよい。
【００１９】
ステップ１１６では、１つまたは複数のクラスタの、クラスタのそれぞれにおいて、式（２）を使用して、「n_i」個のデータサンプルが選択されてもよい。さらに、「n_i」個のデータサンプルを選択する手順は、データセットの１つまたは複数のクラスタのクラスタすべてについて繰り返すことができる。具体的なシナリオでは、クラスタから選択された、ある数のデータサンプルが最小限である場合、プロセス１００は、１つまたは複数のクラスタのそれぞれから、サンプルの数の所定のカウントを選択してもよい。これは、サンプルの数の選択された決定されたカウントが、しきい値未満である場合に行なわれ得る

。ステップ１１８では、合計「ｎ」個のデータサンプルが、１つまたは複数のクラスタのそれぞれから選択され、それにより、データセットのサイズを小さくする。

【００２０】

次に図２を参照すると、本開示のいくつかの実施形態による、１つまたは複数のクラスタからの複数のデータサンプルをソートして選択するためのプロセス２００が図示される。

【００２１】

ステップ２０４では、１つまたは複数のクラスタのうち、１つのクラスタから第１のデータサンプルが選択されてもよい。ステップ２０６では、第１の選択されたサンプルから最も遠い、第２のデータサンプルが選択されてもよい。ステップ２０８では、第１のデータサンプルと第２のサンプルは、データセット内に維持されてもよい。ステップ２１０では、第３のデータサンプルが選択されてもよい。第３のデータサンプルの選択は、ステップ２１６において、あるメカニズムにしたがって実施され、データセットの外部から、例えば第３のデータサンプルから、ランダムなサンプルが選択される。データセットの外部のデータサンプルに関して、データセットのデータサンプルの距離を、決定することができる。例えば、第３のデータサンプルの距離は、データセットの第１のデータサンプルに関して「ｄ１３」として、およびデータセットの第２のデータサンプルに関して「ｄ２３」として決定されてもよい。

【００２２】

距離「ｄ１３」と距離「ｄ２３」の距離のうち、小さいほうを選択することができる。小さいほうの距離とは、例えば図２で図示されるような「ｄ１３」である。さらに、データセットの外部に存在するデータサンプルのすべてについて、距離をチェックして最小距離を選択する、上で言及したステップが決定されてもよい。例えば、データセットの外部の別のデータサンプルは、第４のデータサンプルであってもよく、決定される距離は、例えば、データセットの第１のデータサンプルから第４のデータサンプルまで、「ｄ１４」として、およびデータセットの第２のデータサンプルからの「ｄ２４」としてであってもよい。

【００２３】

続いて、決定される距離の最小を、例えば「ｄ１３」および「ｄ２４」として決定することができる。次に、最小の決定された距離からの最大距離、例えば「ｄ１３」が選択される。最終的に、データサンプル、例えば最大距離に対応する第３のサンプル、例えば「ｄ１３」が選択されてもよく、データセットに挿入されてもよい。

【００２４】

層化サンプリングメカニズムを使用することによって、および、ばらつきの最大化に基づいてクラスタ内でデータサンプルをソートすることによって、ステップ２１２では、「n_i」個のサンプルを、１つまたは複数のクラスタのクラスタから選択することができ、それによって、選択されたサンプルは一意であり、そのクラスタのばらつき全体をカバーすることができる。

【００２５】

ステップ２１４では、上述のステップ２０４～２１２は、１つまたは複数のクラスタの、クラスタのそれぞれについて繰り返され、データセットの性質を保った新しい縮小データセットを作成することができる。

【００２６】

例示の実施形態では、手書きのアルファベットのデータセットがあると仮定する。データセットでは、アルファベット「ａ」は、様々なユーザによって、イタリック体、ボールド体、異なるフォントサイズで、または筆記体など、異なった形で書かれ得る。さらに、アルファベットは、アルファベットが「ａ」、「ｂ」、「ｃ」などのアルファベットのカテゴリに存在するかどうかに基づいてクラスタ化され得る。アルファベット「ａ」のクラスタからの複数のデータサンプルを検討する。アルファベット「ａ」のクラスタでは、イタリック体の「ａ」は、数が少ないものとする。例えば、５０Ｋデータサンプルを有するデータセット中、５Ｋデータサンプルだけが、アルファベット「ａ」のイタリック体を表現

する場合があり、故に５０Ｋデータサンプル内で、イタリック体の「ａ」の一意性を表現し得る。一意なイタリック体「ａ」は、データサンプルを配列してソートするために使用されてもよい。したがって、５０Ｋサンプルから、５Ｋサンプルが、アルファベット「ａ」のイタリック体を表現するために使用することができると結論付けることができる。さらに、これらの表現される５Ｋサンプルのうち、どれが選ばれるかは、クラスタ内でのデータサンプルのばらつきの最大化に基づいたソーティングメカニズムによって決定されてもよい。

【００２７】

次に図３を参照すると、実施形態による、クラスタを表現するためにサンプルを選択する方法３００のフローチャートが図示される。ステップ３０２では、１つまたは複数のクラスタが受信される。１つまたは複数のクラスタのそれぞれは、複数のサンプルを含んでもよい。

【００２８】

ステップ３０４では、１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントが決定されてもよい。１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントは、１つまたは複数のクラスタのそれぞれについてのサイズ、ばらつき、およびクラスタ確率のうちの少なくとも１つに基づいて、層化サンプリング技法を使用して、決定されてもよい。クラスタ確率は、機械学習（ＭＬ）モデルを使用して決定されてもよく、ＭＬモデルは、クラスタの複数のサンプルを分類する。訓練無しのＭＬモデルの事例では、各クラスタは、等しい確率を割振られることに留意されたい。

【００２９】

ステップ３０６では、１つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列が生成されてもよい。例えば、アレイベースの距離行列は、ユークリッド距離行列であってもよい。ステップ３０８では、クラスタの複数のサンプルは、クラスタ内の複数のサンプルのばらつきの度合いに基づいて、１つまたは複数のクラスタのそれぞれについてのアレイベースの距離行列を使用して、ソートすることができる。

【００３０】

ステップ３１０では、クラスタを表現するために、サンプルの数の決定されたカウントを、複数のクラスタのそれぞれの、ソートされた複数のサンプルから選択してもよい。加えて、いくつかの実施形態では、サンプルの数の選択された決定されたカウントがしきい値未満である場合、サンプルの数の所定のカウントは、１つまたは複数のクラスタのそれぞれから選択されてもよい。

【００３１】

１つまたは複数のコンピュータ可読記憶媒体は、本開示と一貫性のある実施形態を実装する際に利用される。コンピュータ可読記憶媒体とは、プロセッサによって可読な情報またはデータを記憶することができる、あらゆるタイプの物理メモリを称する。故に、コンピュータ可読記憶媒体は、本明細書で説明される実施形態と一貫性のあるステップまたは工程をプロセッサに実行させるための命令を含む、１つまたは複数のプロセッサによる実行のための命令を記憶する。「コンピュータ可読媒体」という用語は、有形物を含み、搬送波および一過性の信号を含まない、つまり非一時的であると理解されるべきである。例としては、ランダムアクセスメモリ（ＲＡＭ）、読み取り専用メモリ（ＲＯＭ）、揮発性メモリ、非揮発性メモリ、ハードドライブ、ＣＤ　ＲＯＭ、ＤＶＤ、フラッシュドライブ、ディスク、およびあらゆる他の既知の物理的な記憶媒体が挙げられる。

【００３２】

明確にするために、上述のことは、本開示の実施形態を、異なる機能的なユニットおよびプロセッサに関して説明したものであることを諒解されたい。しかしながら、異なる機能的なユニット、プロセッサ、またはドメイン間での機能性のあらゆる好適な分散が、本開示から逸脱することなく使用され得ることが明らかとなろう。

【００３３】

例えば、別個のプロセッサまたはコントローラによって実施されるよう図示される機能性

は、同一のプロセッサまたはコントローラによって実施されてもよい。したがって、特定の機能的なユニットへの参照は、厳密な論理的または物理的な構造または編成を示すのではなく、説明される機能性を提供するための好適な手段への参照として考えられるに過ぎない。

【００３４】

いくつかの実施形態に関連して本開示を説明したが、本開示は、本明細書で説明した特定の形態に限定されるよう意図されていない。そうではなく、本開示の範囲は、特許請求の範囲によってのみ制限される。加えて、特徴は、特定の実施形態に関連して説明されるように見えるかもしれないが、当業者であれば、説明される実施形態の様々な特徴は、本開示にしたがって組み合わせることができることを認識されよう。

【００３５】

さらには、個々に列挙されているが、複数の手段、要素、またはプロセスのステップは、例えば単一のユニットまたはプロセッサによって実装されてもよい。加えて、個々の特徴が、異なる請求項に含まれる場合があるが、これらは、可能であれば有利に組み合わされてもよく、異なる請求項に含まれることは、特徴の組み合わせが、実行可能ではない、および／または有利ではないということを意味するものではない。また、特徴が、請求項の１つのカテゴリに含まれることは、このカテゴリへの限定することを意味するものではなく、むしろ、適当であれば特徴は他の請求項カテゴリに等しく適用可能であり得る。

【図１】

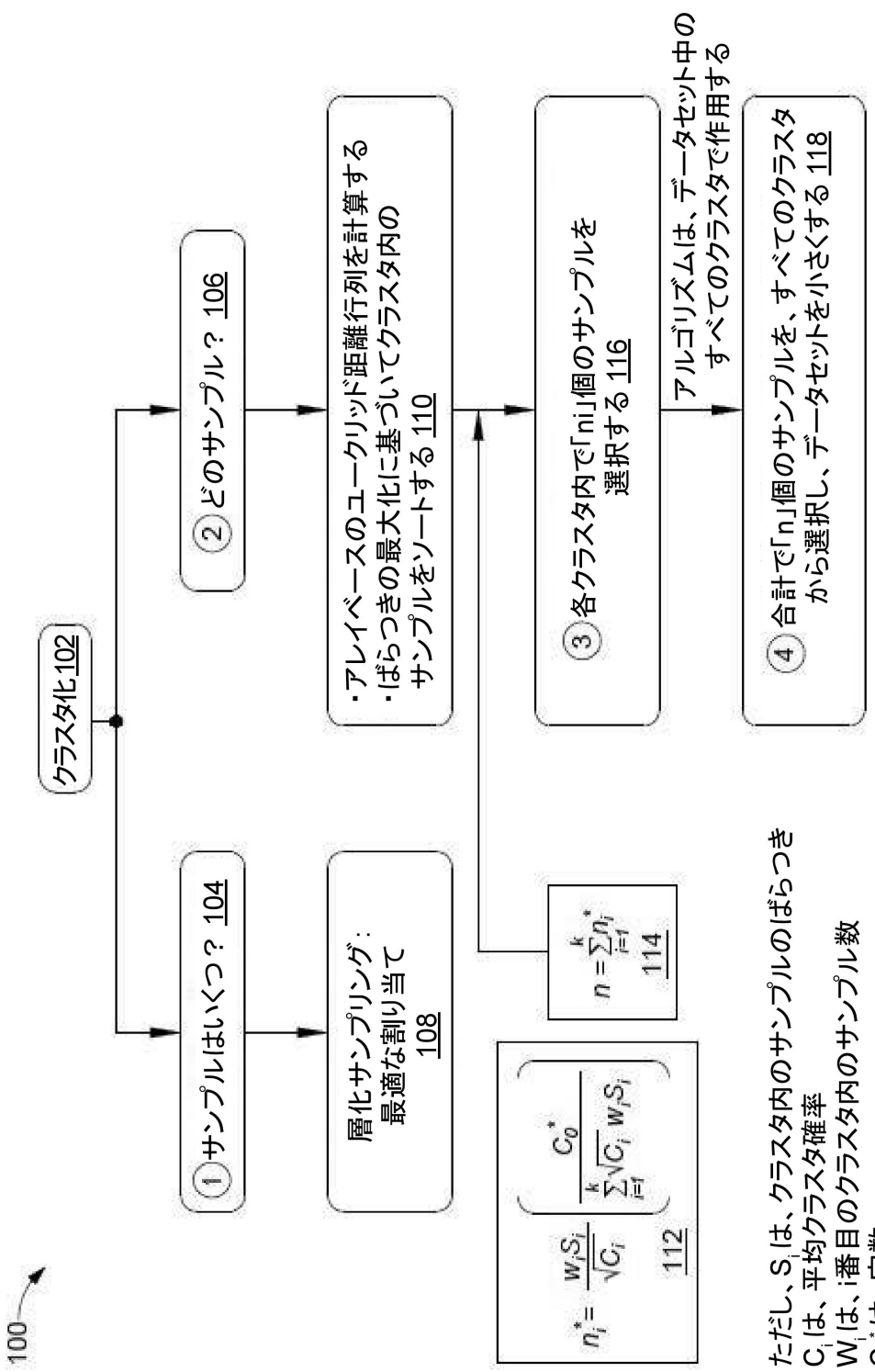
100
クラスタ化102
①サンプルはいくつ？ 104
②どのサンプル？ 106
層化サンプリング：
最適な割り当て
108
・アレイベースのユークリッド距離行列を計算する
・ばらつきの最大化に基づいてクラスタ内の
サンプルをソートする 110
③各クラスタ内で「ni」個のサンプルを
選択する 116
④合計で「n」個のサンプルを、すべてのクラスタ
から選択し、データセットを小さくする 118
アルゴリズムは、データセット中の
すべてのクラスタで作用する
$n_i^* = \frac{w_i S_i}{\sqrt{C_i}} \left(\frac{C_0^*}{\sum_{i=1}^{k} \sqrt{C_i} w_i S_i} \right)$ 112
$n = \sum_{i=1}^{k} n_i^*$ 114
ただし、S_iは、クラスタ内のサンプルのばらつき
C_iは、平均クラスタ確率
W_iは、i番目のクラスタ内のサンプル数
C_o^*は、定数

図１

【図２】

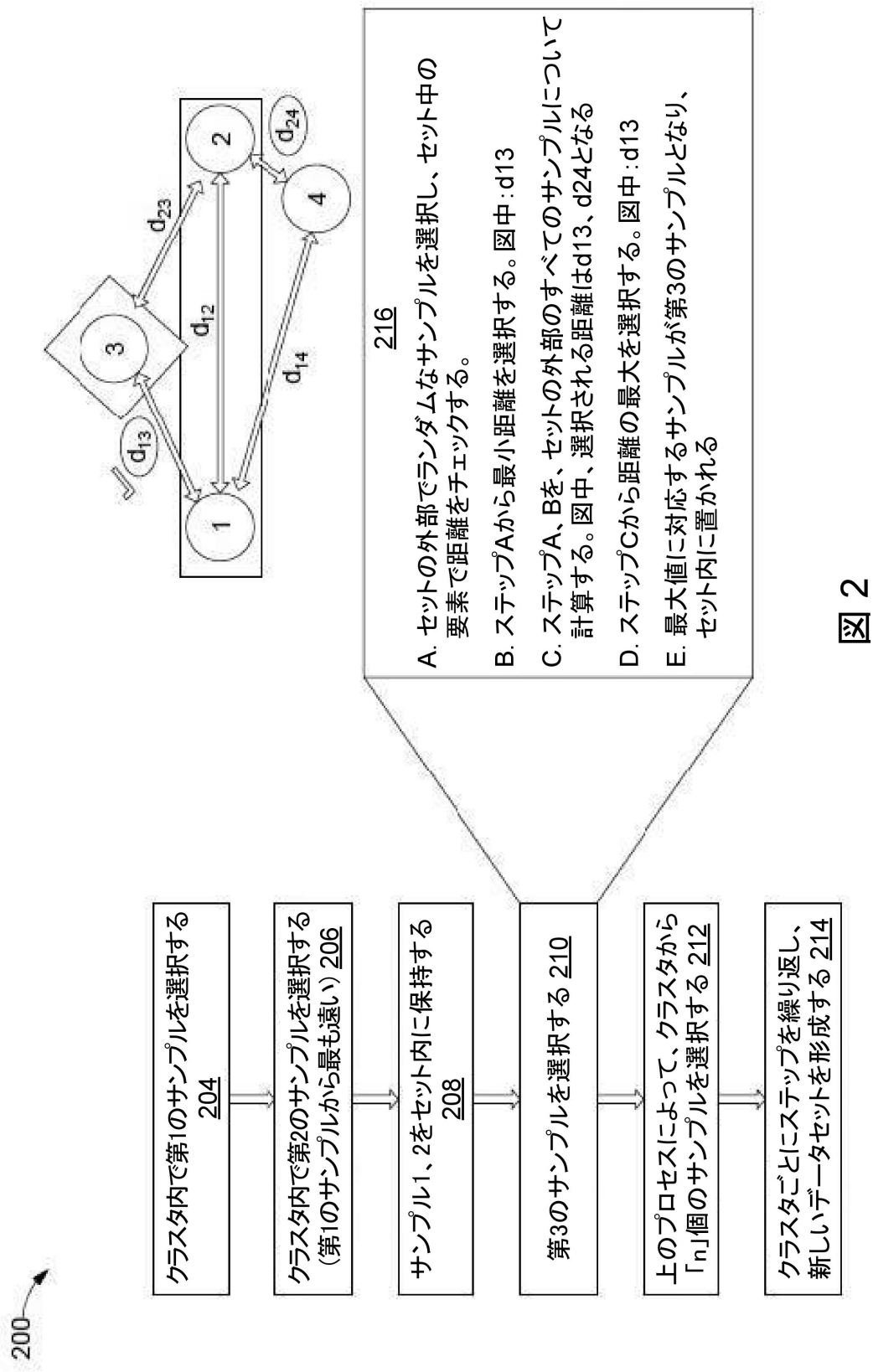
200
クラスタ内で第1のサンプルを選択する 204
クラスタ内で第2のサンプルを選択する（第1のサンプルから最も遠い）206
サンプル1、2をセット内に保持する 208
第3のサンプルを選択する 210
上のプロセスによって、クラスタから「n」個のサンプルを選択する 212
クラスタごとにステップを繰り返し、新しいデータセットを形成する 214
216
A. セットの外部でランダムなサンプルを選択し、セット中の要素で距離をチェックする。
B. ステップAから最小距離を選択する。図中：d13
C. ステップA、Bを、セットの外部のすべてのサンプルについて計算する。図中、選択される距離はd13、d24となる
D. ステップCから距離の最大を選択する。図中：d13
E. 最大値に対応するサンプルが第3のサンプルとなり、セット内に置かれる
1
2
3
4
d_{12}
d_{13}
d_{14}
d_{23}
d_{24}

図２

【図３】

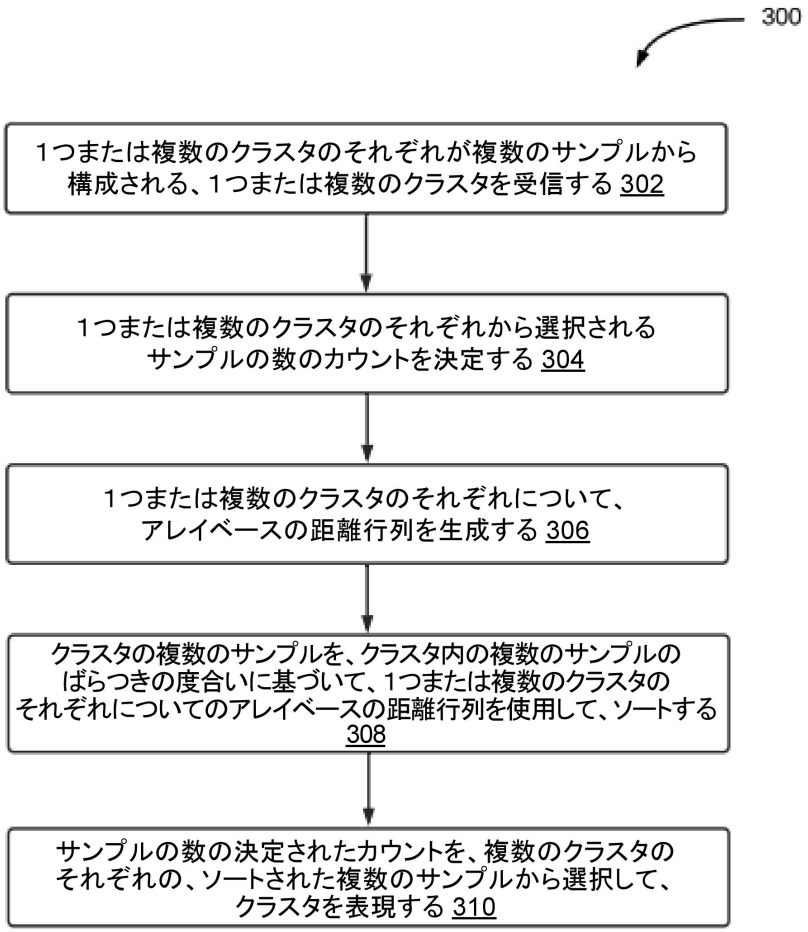
300
1つまたは複数のクラスタのそれぞれが複数のサンプルから構成される、1つまたは複数のクラスタを受信する 302
1つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントを決定する 304
1つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列を生成する 306
クラスタの複数のサンプルを、クラスタ内の複数のサンプルのばらつきの度合いに基づいて、1つまたは複数のクラスタのそれぞれについてのアレイベースの距離行列を使用して、ソートする 308
サンプルの数の決定されたカウントを、複数のクラスタのそれぞれの、ソートされた複数のサンプルから選択して、クラスタを表現する 310

図 3

【手続補正書】
【提出日】令和4年8月16日(2022.8.16)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
クラスタを表現するためにサンプルを選択する方法であって、
最適化デバイスによって、それぞれが複数のサンプルを備える１つまたは複数のクラスタを受信し（３０６）、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれについてのサイズ、ばらつき、およびクラスタ確率のうちの少なくとも１つに基づいて、層化サンプリング技法を使用して、前記１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントを決定し（３０４）、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列を生成し（３０６）、
前記最適化デバイスによって、前記クラスタの前記複数のサンプルを、前記クラスタ内の前記複数のサンプルのばらつきの度合いに基づいて、前記１つまたは複数のクラスタのそれぞれについての前記アレイベースの距離行列を使用して、ソートし（３０８）、
前記最適化デバイスによって、前記クラスタを表現するために、サンプルの数の前記決定されたカウントを（３１０）、前記複数のクラスタのそれぞれの、前記ソートされた複数のサンプルから選択する、
方法。
【請求項２】
前記サンプルの数の前記選択された決定されたカウントがしきい値未満である場合、前記サンプルの数の所定のカウントを、前記１つまたは複数のクラスタのそれぞれから選択し、前記しきい値が、各データセットに特有であり、クラスタのサイズを前記データセットと比較することによって決定される、
請求項１に記載の方法。
【請求項３】
前記クラスタ確率が、機械学習（ＭＬ）モデルを使用して決定され、前記ＭＬモデルが、前記クラスタの前記複数のサンプルを分類する、
請求項１に記載の方法。
【請求項４】
下記を行なうように構成された１つまたは複数のコンピューティングデバイスを備える、
最適化デバイスによって、それぞれが複数のサンプルを備える１つまたは複数のクラスタを受信する、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれについてのサイズ、ばらつき、およびクラスタ確率のうちの少なくとも１つに基づいて、層化サンプリング技法を使用して、前記１つまたは複数のクラスタのそれぞれから選択されるサンプルの数のカウントを決定する、
前記最適化デバイスによって、前記１つまたは複数のクラスタのそれぞれについて、アレイベースの距離行列を生成する、
前記最適化デバイスによって、前記クラスタの前記複数のサンプルを、前記クラスタ内の前記複数のサンプルのばらつきの度合いに基づいて、前記１つまたは複数のクラスタのそれぞれについての前記アレイベースの距離行列を使用して、ソートする、および、
前記最適化デバイスによって、前記クラスタを表現するために、サンプルの数の前記決定されたカウントを、前記複数のクラスタのそれぞれの、前記ソートされた複数のサンプル

から選択する、
システム。
【請求項５】
前記サンプルの数の前記選択された決定されたカウントがしきい値未満である場合、前記サンプルの数の所定のカウントを、前記１つまたは複数のクラスタのそれぞれから選択し、前記しきい値が、各データセットに特有であり、クラスタのサイズを前記データセットと比較することによって決定される、
請求項４に記載のシステム。
【請求項６】
前記クラスタ確率が、機械学習（ＭＬ）モデルを使用して決定され、前記ＭＬモデルが、前記クラスタの前記複数のサンプルを分類する、
請求項４に記載のシステム。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2022/052333

A. CLASSIFICATION OF SUBJECT MATTER
G06N20/00 Version=2022.01

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Databases: IPO Internal Database, Patseer
Keywords: Cluster, sample screening, re-grouping, representation

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US20160180556A1 (DENG CHANG) (23-06-2016) 23 June 2016 Abstract, paragraphs [0003], [0007], [0009]-[0011], [0062]-[0063], [0065]-[0066], figure 4	1-8
X	CN107194430A (BEIJING SANKUAI ON LINE SCIENCE & TECHNOLOGY CO LTD) (22-09-2017) 22 September 2017 Whole Document	1-8

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21-06-2022	21-06-2022
Name and mailing address of the ISA/ Indian Patent Office Plot No.32, Sector 14,Dwarka,New Delhi-110075 Facsimile No.	Authorized officer Shri Ram Kaunaujia Telephone No. +91-1125300200

Form PCT/ISA/210 (second sheet) (July 2019)

フロントページの続き

(81)指定国・地域 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

(72)発明者 シン， マドゥスダン
インド ５６０１００ カルナータカ， バンガロール， エレクトロニック シティー－１，ニーラドリー ロード， ファースト クロス， アジメラ ストーン パーク， ビー－６０３
(72)発明者 バララマン， ムリダル
インド ５６００３７ カルナータカ， バンガロール， ドダネクンディ， チナパナハリ メインロード， エスブイエス パームズ ２， ビー ２０６
(72)発明者 デブナス， スカント
インド ３８２４８１ グジャラート， アフマダーバード， チャンドロディヤ， エヌアール プラカシュ ナガー ソサエティ， パート－３， ガジュラジ ソサエティ， １７７
(72)発明者 グプタ， ムリナル
インド １８０００２ ジャンムー カシミール， ジャンムー， タラブ ティロ， バサント ビハール， ７８－エー

Ｆターム(参考) 5B175 FA03 FB04